MAKING THE RUSSIAN FLORA VISIBLE: FAST DIGITISATION OF THE MOSCOW UNIVERSITY HERBARIUM (MW) IN 2015

Digitisation of collections (databasing of label information and imaging) is a recent trend in herbarium management (Flannery, 2012). Public access to such information via the Internet makes collections more broadly useful and improves scientific research (Smith & Blagoderov, 2012).

Today, at least 61 herbaria have over 1M physical specimens. World leaders in herbarium digitisation are P, L, NY, PE and US (Table 1). This list is a compilation of data published in open sources and may not be complete, but it shows that we are far from full digitisation of the world's largest herbaria. This "digital ranking" is unstable due to varying number and productivity of scanners used in world herbaria depending on budget fluctuation. For instance, NY is currently adding 20K scans per month, K and BM have recently announced a large-scale digitisation project using DigiStreet system based in Amsterdam making 3K scans per day, and US is currently scanning 4K per day of ferns and daisies. Most of digitised specimens are available on the Internet, but quite a few images are stored offline waiting for basic indexing.

The Moscow University Herbarium (MW) received a grant from the Russian Science Foundation (8,000,000 RUR; 142,300 USD) in 2015 for digitisation and over 0.5M specimens were scanned. MW is focused on the flora of temperate Eurasia with a pronounced emphasis on the flora of Russia. With five staff members, it is the second-largest herbarium in Russia after the Komarov Institute (LE). Collections of MW include 989K specimens of 36,289 species of vascular plants and 2001 species of bryophytes. MW holds some important historical collections by G.F. Hoffmann, J.F. Ehrhart, C.B. Trinius, J.R. and J.G.A. Forster, etc. So far we have found 4.5K type specimens.

Approximately 600 type specimens were previously digitized—thanks to financial support from RFBR – and are available online (http://herba.msu.ru/pictures/mw_type/index.html). Sixty-three historical specimens derived from Carl Linnaeus collections were published on CD *Herbarium Linnaeanum* (Balandin & al., 2001).

In 2015, as a result of the Global Plants Initiative, LE digitised ca. 6.1K specimens and became the largest virtual herbarium in Russia. Also, ca. 5K have been scanned in St. Petersburg University Herbarium (LECB) (V.A. Bubyreva, pers. comm.), but they are not available online.

In terms of the 'five task clusters' introduced by Nelson & al. (2012), the fourth and fifth stages of digitisation (label data capture and georeferencing) are not the top priority of this project. Instead we decided to scan all specimens at 300 dpi in 2015–2018, to digitise types at 600 dpi, and finally to database and georeference label data from as many specimens as possible.

Table 1. Number of scanned herbarium specimens in the world's largest virtual herbaria (as of early 2015).

Herbarium code	No. of specimens scanned	URL
P	5.3M	https://science.mnhn.fr/institution/mnhn/collection/p/item/search/form
L	4.0M	https://science.naturalis.nl/en/collection/digitization/digitizing-herbarium/
NY	2.0M	http://www.eurekalert.org/pub_releases/2013-10/tnyb-tma100213.php
PE	1.8M	http://pe.ibcas.ac.cn/en/
US	1.5M	https://emu.kesoftware.com/downloads/EMu/UserGroupMeetings/2014_NA/Presentations/D3_-_2_-_NMNH_-_A_Tale_of_Two_Crowdsourcing_Venues.ppsx
UPS	0.62M	http://www.herbarium-ume.se/virtuella_herbariet/
MEXU	0.50M	http://www.revista.unam.mx/vol.15/num4/art30/index.html
LD	0.45M	http://www.herbarium-ume.se/virtuella_herbariet/
H	0.44M	http://digitarium.fi/content/statistics/
E	0.30M	http://elmer.rbge.org.uk/bgbase/vherb/bgbasevherb.php
K	0.30M	http://www.kew.org/science-conservation/collections/herbarium
GH	0.30M	http://kiki.huh.harvard.edu/databases/specimen_index.html
MO	0.23M	http://www.tropicos.org/ImageSearch.aspx
UC+JEPS	0.20M	https://webapps.cspace.berkeley.edu/ucjeps/publicsearch/publicsearch/
O	0.17M	http://dx.doi.org/10.17161/bi.v9i1.4748
B	0.15M	http://ww2.bgbm.org/herbarium/default.cfm
BM	0.14M	http://data.nhm.ac.uk/

Preparation

Due to the long tradition of floristic research, MW consists of 11 independent regional collections of vascular plants: Eastern Europe, Asian Russia, Caucasus, Crimea, Middle Asia, Mongolia, Western Europe, other Asian countries, Africa, America, Australia & Oceania. The Asian Russia (155K specimens) was the top priority for 2015 followed by Eastern Europe (339K). In 1992–2004, all specimens of MW were counted by species and regions and the records were updated annually, therefore, at the beginning of this project we knew that we had 38,355 taxa of vascular plants (including 36,289 identified species). Before scanning, we checked the taxonomy and basic synonymy. Beginning in April 2015, 5 staff and ca. 60 volunteers have barcoded 600K specimens and performed basic curation procedures.

Technical issues of scanning

Collections were scanned by a commercial partner using three to seven individual planetary A2 scanners with manual placing of specimens. Scanners were operated from late May to September 2015 with an average productivity of ca. 1K specimens per day per scanner. The basic file for each specimen has TIFF format, 300 dpi and ca. 60 MB size. A smaller copy has JPG format, 0.8 compression, 300 dpi and ca. 2.5 MB size. Several stages of computer processing without operator's manual labor included: 90° rotation, cosmetic rotation to a right angle, cutting of a black background, barcode reading and file renaming, and JPG copy production.

Metadata for each specimen were generated by a scanner operator, who searched for exact matches between species name and area code on a folder and in the database. After that all specimens from the folder received the same metadata. So, one by one all specimens were captured and indexed in the database with four fields: generic 4-digit code, taxon name, one of 30 area codes (21 in Eastern Europe and 9 in Asian Russia), barcode (ID). An example is below:

2195_Rumex_acetosella / E1 / MW0319190

Key results

In total 502K specimens from Asian Russia (156K) and Eastern Europe (346K) were digitised in 19 weeks (6 days a week), including 76K specimens from Arctic and 24K specimens from Russian Far East. In 2016, 88K specimens from Russian Caucasus and the Crimea will be digitised to make the Russian virtual collections comprehensive. Currently, we hold the largest digitised herbarium collections from Russia, Belarus, Ukraine, Moldova,

Moscow University Herbarium (MW) scanner being operated by Maxim Belyakov. Each scanner has a production of ca. 1000 specimens per day. — Photo by A. Seregin.

Kazakhstan, and likely from Lithuania, Latvia, and Estonia. At the moment, we operate the eighth-largest digital herbarium in the world (yet offline).

Moscow University spent 7,000,000 RUR to scan 502K specimens. It is roughly 0.2 USD per specimen as of December 12, 2015 or twice as expensive as in the large-scale scanning of Paris herbarium. On the other hand, a single image processing was considerably cheaper than in NY or L. We regard the productivity of our commercial partner to be satisfactory in the terms of time/money ratio. Keeping in mind that we employed seven scanners, 1K images per day per scanner (300 dpi) in MW could be compared with 689 images (450 dpi) on a single automated digitising line in Helsinki (Tegelberg & al., 2014), 4K to 12K (300 dpi) on two lines in Paris (Pignal & Michiels, 2011), or 20K (300 dpi) on an industrial scale scanning conveyor belt in Leiden which require more staff time (Heerlien & al., 2013).

In November 2015, a number of taxonomists who were actively working with various vascular plant groups from Asian Russia in the last decade, were asked via email to identify plants which were named only to genus or family. Forty-one botanists (77%) responded—22 of them sent identification lists for 422 specimens. It was much cheaper to produce and distribute scans rather than to facilitate visits of 22 botanists from 12 cities and three countries in one of the world's most expensive cities.

Further scanning and label capturing

In 2016, we plan to finish barcoding all 989K specimens and numerous new accessions. Ca. 200K of vascular plant specimens and 80K labels from bryophyte envelopes will be scanned. The target regions for 2016 are Caucasus, Crimea, Mongolia, Southern and Eastern Asia, as well as new accessions to branches digitised a year ago.

Currently, Moscow University does not have an efficient web-portal that allows users to search and study MW specimen images, but we hope that it will be launched soon. For some time, there will be no captured label data in the database, so search queries will be limited by taxa names and codes of curatorial areas, which are fortunately fairly small (250,000 km² on average for Eastern Europe). The web-portal will allow us to gather updated identifications and nomenclature. Re-identified specimens will be rescanned. Small archives of up to 50 JPG files are already available on demand for all researchers.

In 2016, we are planning to start capturing label information. Our hope is that label data from scanned images will be captured by local naturalists. We plan to pay a modest remuneration for professional botanists from small institutions and universities to database labels that are of interest to them.

Acknowledgments

I would like to thank the following taxonomists for quick identification by scans of unnamed specimens: K.S. Baikov, A.A. Bobrov, I.O. Buzunova, V.V. Byalt, V.M. Doronkin, I.V. Enustschenko, N. Friesen, I.G. Gavrilenko, D.A. German, I.I. Gureyeva, I.V. Han, A.A. Keczaykin, M.S. Knjazev, G.Yu. Konechnaja, N.K. Kovtonyuk, G.A. Lazkov, D.G. Melnikov, M.A. Mikhailova, M.V. Olonova, A.I. Shmakov, N.N. Tupitsyna, V.V. Zuev. MW digitisation is supported by the grant "Scientific basis of the national biobank—depository of the living systems " (#14-50-00029) from Russian Science Foundation (RNF).

Literature cited

Balandin, S.A., Gubanov, I.A., Jarvis, C.E., Majorov, S.R., Simonov, S.S., Sokoloff, D.D. & Sukhov, S.V. 2001. *Herbarium Linnaeanum: The Linnaean collection of the Herbarium of Moscow State University: Digital images, comments, historical review.* Moscow: Dehlia. [CD-ROM + booklet]

Flannery, M.C. 2012. Flatter than a pancake: Why scanning herbarium sheets shouldn't make them disappear. *Spontaneous Generations: A Journal for the History and Philosophy of Science* 6(1): 225–232. http://dx.doi.org/10.4245/sponge.v6i1.16134

Heerlien, M., Van Leusen, J., Schnörr, S. & Van Hulsen, K. 2013. The natural history production line. Pp. 289–294 in: Addison, A.C., Guidi, G., De Luca, L. & Pescarin, S. (eds.), *2013 Digital Heritage International Congress (DigitalHeritage)*, 28 Oct–1 Nov 2013, Marseille, France, vol. 2. Piscataway: IEEE. http://dx.doi.org/10.1109/DigitalHeritage.2013.6744766

Nelson, G., Paul, D., Riccardi, G. & Mast, A.R. 2012. Five task clusters that enable efficient and effective digitization of biological collections. *ZooKeys* 209: 19–45. http://dx.doi.org/10.3897/zookeys.209.3135

Pignal, M. & Michiels, H. 2012. Switching to the fast track: Rapid digitization of the world's largest herbarium. Botany 2011 – Colombus, Ohio. 11 Jul. 2012. Presentation available at: http://collections.mnhn.fr/wiki/attach/Visit_October2012/Paris-Herbarium-Digitization_2012-07-12.pdf (accessed 14 Dec 2015).

Smith, V.S. & Blagoderov, V. 2012. Bringing collections out of the dark. *ZooKeys* 209, 1–6. http://dx.doi.org/10.3897/zookeys.209.3699

Tegelberg, R., Mononen, T. & Saarenmaa, H. 2014. High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. *Taxon* 63: 1307–1313. http://dx.doi.org/10.12705/636.13

Alexey P. Seregin
Herbarium, Lomonosov Moscow State University, Moscow, 119991, Russia; botanik.seregin@gmail.com